Plots, Prediction and Testing for the Multivariate Linear Model

David J. Olive*

Southern Illinois University

December 18, 2013

Abstract

To use the multivariate linear model, reliable prediction regions and tests of hypotheses are needed. This paper provides useful prediction regions, gives F approximations to the widely used Wilks' Λ, Pillai's trace, and Hotelling Lawley trace test statistics, and gives plots to check goodness and lack of fit, to check for outliers and influential cases, and to check whether the error distribution is multivariate normal or from some other elliptically contoured distribution. Some of the plots and prediction regions can be extended to more general multivariate regression models.

KEY WORDS: MANOVA; multivariate linear regression; multivariate regression; prediction regions; seemingly unrelated regressions.

*David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. E-mail address: dolive@siu.edu.

1 INTRODUCTION

For the multivariate linear model, this paper provides useful prediction regions and plots, and gives F approximations to the widely used Wilks' Λ, Pillai's trace, and Hotelling Lawley trace test statistics. First notation is given for the multivariate linear model.

The *multivariate linear model* $\boldsymbol{y}_i = \boldsymbol{B}^T\boldsymbol{x}_i + \boldsymbol{\epsilon}_i$ for $i = 1, ..., n$ has $m \geq 2$ response variables $Y_1, ..., Y_m$ and p predictor variables $x_1, x_2, ..., x_p$. Multivariate linear regression and MANOVA models are special cases. The ith case is $(\boldsymbol{x}_i^T, \boldsymbol{y}_i^T) = (x_{i1}, x_{i2}, ..., x_{ip}, Y_{i1}, ..., Y_{im})$. If a constant $x_{i1} = 1$ is in the model, then x_{i1} could be omitted from the case. The model is written in matrix form as $\boldsymbol{Z} = \boldsymbol{XB} + \boldsymbol{E}$ where the matrices are defined below. The model has $E(\boldsymbol{\epsilon}_k) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}_k) = \boldsymbol{\Sigma_\epsilon} = (\sigma_{ij})$ for $k = 1, ..., n$. Also $E(\boldsymbol{e}_i) = \boldsymbol{0}$ while $\mathrm{Cov}(\boldsymbol{e}_i, \boldsymbol{e}_j) = \sigma_{ij}\boldsymbol{I}_n$ for $i, j = 1, ..., m$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $\boldsymbol{e}_i$ is defined below. Then the $p \times m$ coefficient matrix $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{\beta}_1 & \boldsymbol{\beta}_2 & \dots & \boldsymbol{\beta}_m \end{bmatrix}$ and the $m \times m$ covariance matrix $\boldsymbol{\Sigma_\epsilon}$ are to be estimated, and $E(\boldsymbol{Z}) = \boldsymbol{XB}$ while $E(Y_{ij}) = \boldsymbol{x}_i^T\boldsymbol{\beta}_j$.

The $n \times m$ matrix of response variables and $n \times m$ matrix of errors are

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Y}_1 & \boldsymbol{Y}_2 & \dots & \boldsymbol{Y}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^T \\ \vdots \\ \boldsymbol{y}_n^T \end{bmatrix} \text{ and } \boldsymbol{E} = \begin{bmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \dots & \boldsymbol{e}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_1^T \\ \vdots \\ \boldsymbol{\epsilon}_n^T \end{bmatrix},$$

while the $n \times p$ design matrix of predictor variables is $\boldsymbol{X}$.

Least squares is the classical method for fitting the multivariate linear model. The *least squares estimators* are $\hat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 & \hat{\boldsymbol{\beta}}_2 & \dots & \hat{\boldsymbol{\beta}}_m \end{bmatrix}$. The matrix of *predicted values* or *fitted values* $\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{B}} = \begin{bmatrix} \hat{\boldsymbol{Y}}_1 & \hat{\boldsymbol{Y}}_2 & \dots & \hat{\boldsymbol{Y}}_m \end{bmatrix}$. The matrix of *residuals* $\hat{\boldsymbol{E}} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = \boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}} = \begin{bmatrix} \hat{\boldsymbol{r}}_1 & \hat{\boldsymbol{r}}_2 & \dots & \hat{\boldsymbol{r}}_m \end{bmatrix}$. These quantities can be found

from the m multiple linear regressions of Y_j on the predictors: $\widehat{\boldsymbol{\beta}}_j = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}_j$, $\widehat{\boldsymbol{Y}}_j = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_j$ and $\hat{\boldsymbol{r}}_j = \boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j$ for $j = 1, ..., m$. Hence $\hat{\epsilon}_{i,j} = Y_{i,j} - \hat{Y}_{i,j}$ where $\hat{\boldsymbol{Y}}_j = (\hat{Y}_{1,j}, ..., \hat{Y}_{n,j})^T$. Finally,

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d} = \frac{(\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}})}{n-d} = \frac{(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})^T(\boldsymbol{Z} - \boldsymbol{X}\hat{\boldsymbol{B}})}{n-d} = \frac{\hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}}}{n-d} = \frac{1}{n-d}\sum_{i=1}^{n}\hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i^T.$$

The choices $d = 0$ and $d = p$ are common. If $d = 1$, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon},d=1} = \boldsymbol{S}_r$, the sample covariance matrix of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ since the sample mean of the $\hat{\boldsymbol{\epsilon}}_i$ is $\boldsymbol{0}$. Let $\hat{\boldsymbol{\Sigma}}_\epsilon = \hat{\boldsymbol{\Sigma}}_{\epsilon,p}$ be the unbiased estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$.

The $\boldsymbol{\epsilon}_i$ are assumed to be iid. Some important joint distributions for $\boldsymbol{\epsilon}$ are completely specified by an $m \times 1$ population *location* vector $\boldsymbol{\mu}$ and an $m \times m$ symmetric positive definite population *dispersion* matrix $\boldsymbol{\Sigma}$. An important model is the elliptically contoured $EC_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution with probability density function

$$f(\boldsymbol{z}) = k_m|\boldsymbol{\Sigma}|^{-1/2}g[(\boldsymbol{z} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z} - \boldsymbol{\mu})]$$

where $k_m > 0$ is some constant and g is some known function. The multivariate normal (MVN) $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution is a special case.

Some additional notation will be useful. Assume that $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_n$ are iid from a multivariate distribution. The classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ of multivariate location and dispersion is the sample mean and sample covariance matrix where

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \quad \text{and} \quad \boldsymbol{S} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\boldsymbol{x}_\mathrm{i} - \overline{\boldsymbol{x}})(\boldsymbol{x}_\mathrm{i} - \overline{\boldsymbol{x}})^\mathrm{T}. \tag{1}$$

Let the $p \times 1$ column vector T be a multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $\boldsymbol{C}$ be a dispersion estimator. Then the ith *squared*

sample Mahalanobis distance is the scalar

$$D_i^2 = D_i^2(T, \boldsymbol{C}) = (\boldsymbol{x}_i - T)^T \boldsymbol{C}^{-1} (\boldsymbol{x}_i - T) \tag{2}$$

for each observation $\boldsymbol{x}_i$. Notice that the Euclidean distance of $\boldsymbol{x}_i$ from the estimate of center T is $D_i(T, \boldsymbol{I}_p)$. The classical Mahalanobis distance uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. Following Johnson (1987, pp. 107-108), the population squared Mahalanobis distance

$$U \equiv D_{\boldsymbol{x}}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}), \tag{3}$$

and for elliptically contoured distributions, U has probability density function (pdf)

$$h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p u^{p/2-1} g(u). \tag{4}$$

The classical large sample $100(1-\delta)\%$ prediction region for a future value $\boldsymbol{x}_f$ given iid data $\boldsymbol{x}_1, ..., , \boldsymbol{x}_n$ is $\{\boldsymbol{x} : D_{\boldsymbol{x}}^2(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq \chi_{p,1-\delta}^2\}$, while for multivariate linear regression, the classical large sample $100(1-\delta)\%$ prediction region for a future value $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ and past data $(\boldsymbol{x}_1, \boldsymbol{y}_i), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$ is $\{\boldsymbol{y} : D_{\boldsymbol{y}}^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq \chi_{m,1-\delta}^2\}$. See Johnson and Wichern (1988, pp. 134, 151, 312). By (4), these regions may work for multivariate normal data, but otherwise tend to have undercoverage. Olive (2013) replaced $\chi_{p,1-\delta}^2$ by the order statistic $D_{(U_n)}^2$ where U_n decreases to $\lceil n(1-\delta) \rceil$. Section 3 will use a similar technique to develop possibly the first practical large sample prediction region for the multivariate linear model with unknown error distribution.

Following regression graphics notation, suppress the subscript indicating case for the response variable and vector of predictor variables. Suppose $m = 1$ and the response variable Y is conditionally independent of the vector of predictors $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$, where the d-dimensional function $h : \mathcal{R}^p \to \mathcal{R}^d$. The *estimated*

sufficient predictor ESP $= \hat{h}(\boldsymbol{x})$. If $d = 1$, then a *response plot* is a plot of ESP versus Y. Cook and Weisberg (1999a, p. 411) define a *sufficient summary plot* to be a plot that contains all of the sample regression information about the conditional distribution of the response given the predictors. Hence a plot of $h(\boldsymbol{x})$ versus Y is a sufficient summary plot while a response plot of $\hat{h}(\boldsymbol{x})$ versus Y is an estimated sufficient summary plot. As an analogy, a sufficient statistic contains all of the sample information for estimating a parameter θ, and want the dimension of the sufficient statistic as small as possible. Similarly want d as small as possible for the greatest amount of dimension reduction.

The Rousseeuw and Van Driessen (1999) DD plot is a plot of classical Mahalanobis distances versus robust Mahalanobis distances. Simulations and results from Olive (2002) suggest that as $n \to \infty$, the plotted points in the DD plot of the least squares residual vectors $\hat{\boldsymbol{\epsilon}}_i$ will cluster tightly about the identity line with unit slope and zero intercept if the $\boldsymbol{\epsilon}_i$ are iid from a multivariate normal $N_m(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ distribution, and cluster tightly about some line through the origin with slope greater than one for a large class of elliptically contoured distributions.

Section 2 suggests making the m response plots, the m residual plots and the DD plot of the residual vectors. For the multivariate linear model, these plots form the minimal collection of 2-dimensional plots for visualizing the conditional distribution of $\boldsymbol{y}|\boldsymbol{x}$. For this model, $\hat{h}_j(\boldsymbol{x}) = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}_j = \hat{Y}_j$ for $j = 1, ..., m = d$. Since Y_j depends on $\boldsymbol{x}$ through $\boldsymbol{x}^T\boldsymbol{\beta}_j$, the conditional distribution of $Y_j|\boldsymbol{x}$ can be visualized with the response plot of $\boldsymbol{x}^T\hat{\boldsymbol{\beta}}_j$ versus Y_j. The response and residual plots will be used to check the multivariate linear model for linearity, influential cases, and outliers. The DD plot of the residuals is used to check the error distribution, to check for outliers, and to display the prediction

region developed in section 3.

Kakizawa (2009) examines testing for the multivariate linear regression model, showing that the Wilks, Pillai, and Hotelling Lawley test statistics perform well asymptotically for a large class of zero mean error distributions. Section 4 reviews these results and shows that the Hotelling Lawley test statistic is closely related to the partial F statistic for multiple linear regression. Section 5 gives some examples and simulations.

2 Plots for the Multivariate Linear Model

This section suggests using residual plots, response plots, and the DD plot to examine the multivariate linear model. The residual plots are often used to check for lack of fit of the multivariate linear model. The response plots are used to check linearity and to detect influential cases and outliers. The response and residual plots are used exactly as in the $m = 1$ case corresponding to multiple linear regression and experimental design models. See Olive and Hawkins (2005) and Cook and Weisberg (1999a, p. 432; 1999b)

The DD plot of the residual vectors $\hat{\boldsymbol{\epsilon}}_i$ is used to check the error distribution, to detect outliers and to display the nonparametric prediction region developed in section 3. The DD plot suggests that the error distribution is elliptically contoured if the plotted points cluster tightly about a line through the origin as $n \to \infty$. The plot suggests that the error distribution is multivariate normal if the line is the identity line. If n is large and the plotted points do not cluster tightly about a line through the origin, then the error distribution may not be elliptically contoured. Make a DD plot of the continuous predictor variables to check for $\boldsymbol{x}$-outliers. These applications of the DD plot for iid

multivariate data are discussed in Olive (2002, 2013). See Examples 1 and 2.

Make the m response and residual plots for the multivariate linear model. A *response plot* for the jth response variable is a plot of the fitted values $\widehat{Y}_{ij}$ versus the response Y_{ij} where $i = 1, ..., n$. The identity line is added to the plot as a visual aid. A *residual plot* corresponding to the jth response variable is a plot of $\hat{Y}_{ij}$ versus r_{ij}. In a response plot, the vertical deviations from the identity line are the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij}$. Assume that the $\widehat{Y}_{ij}$ take on many values (this assumption is not met by the one way MANOVA model, but is met by the multivariate linear regression model and by some MANOVA models). Suppose the model is good, the error distribution is not highly skewed, and $n \geq 10p$. Then the plotted points should cluster about the identity line in each of the m response plots. If outliers are present or if the plot is not linear, then the current model or data need to be changed or corrected. See Example 1 in section 5.

The response and residual plots for the one way MANOVA model need some notation, and it is useful to use three subscripts. Suppose there are independent random samples from p different populations (treatments), and n_i cases are randomly assigned to p treatment groups with $n = \sum_{i=1}^{p} n_i$. Assume that m response variables $\boldsymbol{y}_{ij} = (Y_{ij1}, ..., Y_{ijm})^T$ are measured for the ith treatment. Hence $i = 1, ..., p$ and $j = 1, ..., n_i$. The Y_{ijk} follow different one way ANOVA models for $k = 1, ..., m$. Assume $E(\boldsymbol{y}_{ij}) = \boldsymbol{\mu}_i = (\mu_{i1}, ..., \mu_{im})^T$ and $\mathrm{Cov}(\boldsymbol{y}_{ij}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. Hence the p treatments have possibly different mean vectors $\boldsymbol{\mu}_i$, but common covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$.

Then for the kth response variable, the *response plot* is a plot of $\hat{Y}_{ijk} \equiv \hat{\mu}_{iok}$ versus Y_{ijk} and the *residual plot* is a plot of $\hat{Y}_{ijk} \equiv \hat{\mu}_{iok}$ versus r_{ijk} where the $\hat{\mu}_{iok}$ are the sample means of the p treatments for the kth response variable. Add the identity line to the

response plot and $r = 0$ line to the residual plot as visual aids. The points in the response plot scatter about the identity line and the points in the residual plot scatter about the $r = 0$ line. The response plot consists of p dot plots, one for each value of $\hat{\mu}_{iok}$. The dot plot corresponding to $\hat{\mu}_{iok}$ is the dot plot of $Y_{i,1,k}, ..., Y_{i,n_i,k}$. Similarly, the residual plot consists of p dot plots, and the plot corresponding to $\hat{\mu}_{iok}$ is the dot plot of $r_{i,1,k}, ..., r_{i,n_i,k}$. Assuming the $n_i \geq 10$, the p dot plots should have roughly the same amount of spread in both the response and residual plots. Again m response and residual plots are made, one for each response variable. See Example 2 in section 5.

3 Asymptotically Optimal Prediction Regions

In this section, we will consider a more general multivariate regression model, and then consider the multivariate linear model as a special case. Given n cases of training or past data $(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$ and a vector of predictors $\boldsymbol{x}_f$, suppose it is desired to predict a future vector $\boldsymbol{y}_f$. Then a large sample $(1 - \delta)100\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{y}_f \in \mathcal{A}_n) \rightarrow 1 - \delta$ as $n \rightarrow \infty$, and is asymptotically optimal if the volume of the region converges in probability to the volume of the population minimum volume covering region. The following technical theorem will be needed to prove Theorem 2, and the proof is in the appendix.

Theorem 1. Let $a_j > 0$ and assume that $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ are consistent estimators of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ for $j = 1, 2$.

a) $D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = o_P(1)$.

b) Let $0 < \delta \leq 0.5$. If $(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - (\boldsymbol{\mu}, a_j\boldsymbol{\Sigma}) = O_p(n^{-\delta})$ and $a_j\hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1} = O_P(n^{-\delta})$,

then

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) - \frac{1}{a_j} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = O_P(n^{-\delta}).$$

Now suppose a prediction region for an $m \times 1$ random vector $\boldsymbol{y}_f$ given a vector of predictors $\boldsymbol{x}_f$ is desired for the multivariate linear model. If we had many cases $\boldsymbol{z}_i = \boldsymbol{B}^T \boldsymbol{x}_f + \boldsymbol{\epsilon}_i$, then we could make a prediction region for $\boldsymbol{z}_i$ using a multivariate location and dispersion model prediction region for m variables. Instead, Theorem 2 will use a prediction region on the pseudodata $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{B}}^T \boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ for $i = 1, ..., n$. This takes the data cloud of the n residual vectors $\hat{\boldsymbol{\epsilon}}_i$ and centers the cloud at $\hat{\boldsymbol{y}}_f$. Note that $\hat{\boldsymbol{z}}_i = (\boldsymbol{B} - \boldsymbol{B} + \hat{\boldsymbol{B}})^T \boldsymbol{x}_f + (\boldsymbol{\epsilon}_i - \boldsymbol{\epsilon}_i + \hat{\boldsymbol{\epsilon}}_i) = \boldsymbol{z}_i + (\hat{\boldsymbol{B}} - \boldsymbol{B})^T \boldsymbol{x}_f + \hat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i = \boldsymbol{z}_i + (\hat{\boldsymbol{B}} - \boldsymbol{B})^T \boldsymbol{x}_f - (\hat{\boldsymbol{B}} - \boldsymbol{B})^T \boldsymbol{x}_i = \boldsymbol{z}_i + O_P(n^{-1/2})$.

If the $\boldsymbol{\epsilon}_i$ are iid from an $EC_m(\boldsymbol{0}, \boldsymbol{\Sigma}, g)$ distribution with continuous decreasing g and nonsingular covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = c\boldsymbol{\Sigma}$ for some constant $c > 0$, then the population asymptotically optimal prediction region is $\{\boldsymbol{y} : D_{\boldsymbol{y}}(\boldsymbol{B}^T \boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}\}$ where $P(D_{\boldsymbol{y}}(\boldsymbol{B}^T \boldsymbol{x}_f, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}) = 1 - \delta$. For example, if the iid $\boldsymbol{\epsilon}_i \sim N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then $D_{1-\delta} = \sqrt{\chi^2_{m,1-\delta}}$. If the error distribution is not elliptically contoured, then the above region still has $100(1-\delta)\%$ coverage, but prediction regions with smaller volume may exist.

A natural way to make a large sample prediction region is to estimate the target population minimum volume covering region, but for moderate samples and many error distributions, the natural estimator that covers $\lceil n(1-\delta) \rceil$ of the cases tends to have undercoverage as high as $min(0.05, \delta/2)$. This empirical result is not too surprising since it is well known that the performance of a prediction region on the training data is

superior to the performance on future data, due in part to the unknown variability of the estimator. To compensate for the undercoverage, let q_n be as in Theorem 2.

Theorem 2. Suppose $\boldsymbol{y}_i = E(\boldsymbol{y}_i) + \boldsymbol{\epsilon}_i = \hat{\boldsymbol{y}}_i + \hat{\boldsymbol{\epsilon}}_i$ where $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, and where the zero mean $\boldsymbol{\epsilon}_f$ and the $\boldsymbol{\epsilon}_i$ are iid for $i = 1, ..., n$. Given $\boldsymbol{x}_f$, suppose the fitted model produces $\hat{\boldsymbol{y}}_f$ and nonsingular $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Let $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ and

$$D_i^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{y}}_f)$$

for $i = 1, ..., n$. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + m/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta m/n), \quad \text{otherwise.}$$

If $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let $0 < \delta < 1$ and $h = D_{(U_n)}$ where $D_{(U_n)}$ is the q_nth sample quantile of the D_i. Let the nominal $100(1 - \delta)\%$ prediction region for $\boldsymbol{y}_f$ be given by $\{\boldsymbol{z} : (\boldsymbol{z} - \hat{\boldsymbol{y}}_f)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{z} - \hat{\boldsymbol{y}}_f) \leq D_{(U_n)}^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}}^2(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}^2\} =$

$$\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}\}. \tag{5}$$

a) Consider the n prediction regions for the data where $(\boldsymbol{y}_{f,i}, \boldsymbol{x}_{f,i}) = (\boldsymbol{y}_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. If the order statistic $D_{(U_n)}$ is unique, then U_n of the n prediction regions contain $\boldsymbol{y}_i$ where $U_n/n \to 1 - \delta$ as $n \to \infty$.

b) If $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then (5) is a large sample $100(1 - \delta)\%$ prediction region for $\boldsymbol{y}_f$.

c) If $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is a consistent estimator of $(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and the $\boldsymbol{\epsilon}_i$ come from an elliptically contoured distribution such that the highest density region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}\}$, then the prediction region (5) is asymptotically optimal.

The proof of Theorem 2 is in the appendix. Notice that for the data $(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$, if $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}$ exists, then $100q_n\%$ of the n cases are in their corresponding prediction region, and $q_n \to 1-\delta$ even if $(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is not a good estimator. Hence the coverage q_n of the training data is robust to model assumptions. Of course the volume of the prediction region could be large if a poor estimator $(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$ is used or if the $\boldsymbol{\epsilon}_i$ do not come from an elliptically contoured distribution. The response, residual and DD plots can be used to check model assumptions.

Also want $n \geq 10\max(p, m)$ for least squares estimators. If n is too small, then multivariate data is sparse and the covering ellipsoid for the training data may be far too small for future data, resulting in severe undercoverage. Coverage can also be arbitrarily bad if there is extrapolation or if $(\boldsymbol{x}_f, \boldsymbol{y}_f)$ comes from a different population than that of the data. Also notice that $q_n = 1-\delta/2$ or $q_n = 1-\delta+0.05$ for $n \leq 20p$. At the training data, the coverage $q_n \geq 1-\delta$, and q_n converges to the nominal coverage $1-\delta$ as $n \to \infty$. Suppose $n \leq 20p$. Then the nominal 95% prediction region uses $q_n = 0.975$ while the nominal 50% prediction region uses $q_n = 0.55$. Prediction distributions depend both on the error distribution and on the variability of the estimator $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$. This variability is typically unknown but converges to 0 as $n \to \infty$. Also, residuals tend to underestimate errors for small n. For moderate n, ignoring estimator variability and using $q_n = 1-\delta$ resulted in undercoverage as high as $\min(0.05, \delta/2)$. Letting the "coverage" q_n decrease to the nominal coverage $1-\delta$ inflates the volume of the prediction region for small n, compensating for the unknown variability of $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})$.

Prediction region (5) can be used for multivariate linear regression using least squares or the Su and Cook (2012) inner envelopes estimator. The region can also be used for

the seemingly unrelated regressions model. Theorem 3 relates prediction region (5) to a prediction region originally created for iid multivariate data.

Theorem 3. For multivariate linear regression, when least squares is used to compute $\hat{\boldsymbol{y}}_f$, $\boldsymbol{S}_r$ and the pseudodata $\hat{\boldsymbol{z}}_i$, prediction region (5) is the Olive (2013) nonparametric prediction region applied to the $\hat{\boldsymbol{z}}_i$.

The proof of Theorem 3 is in the appendix. Olive (2013) derived prediction regions for a future observation $\boldsymbol{x}_f$ given n iid $p \times 1$ random vectors $\boldsymbol{x}_i$. Suppose $(T, \boldsymbol{C})$ is an estimator of multivariate location and dispersion $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ such as the classical estimator $(\overline{\boldsymbol{x}}, \boldsymbol{S})$ given by (1). For $h > 0$, consider the hyperellipsoid

$$\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}. \tag{6}$$

A future observation $\boldsymbol{x}_f$ is in region (6) if $D_{\boldsymbol{x}_f} \leq h$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then (6) is a large sample $(1-\delta)100\%$ prediction region if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the q_nth sample quantile of the D_i with p replacing m. The classical parametric multivariate normal large sample prediction region uses $D_{\boldsymbol{x}_f} \equiv MD_{\boldsymbol{x}_f} \leq \sqrt{\chi^2_{p,1-\delta}}$.

The Olive and Hawkins (2010) RMVN estimator $(T_{RMVN}, \boldsymbol{C}_{RMVN})$ is an easily computed $\sqrt{n}$ consistent estimator of $(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$ under regularity conditions (E1) that include a large class of elliptically contoured distributions, and $c = 1$ for the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. The RMVN estimator also gives useful estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ data even when certain types of outliers are present, and will be the robust estimator used in the DD plots. Also see Zhang, Olive and Ye (2012).

The nonparametric region uses the classical estimator $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$ and $h = D_{(U_n)}$. The semiparametric region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h = D_{(U_n)}$.

The parametric MVN region uses $(T, \boldsymbol{C}) = (T_{RMVN}, \boldsymbol{C}_{RMVN})$ and $h^2 = \chi^2_{p,q_n}$ where $P(W \leq \chi^2_{p,q_n}) = q_n$ if $W \sim \chi^2_p$. All three regions are asymptotically optimal for $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distributions with nonsingular $\boldsymbol{\Sigma}$. Olive (2013) shows that the first two regions are asymptotically optimal under (E1). For distributions with nonsingular covariance matrix $c_X\boldsymbol{\Sigma}$, the nonparametric region is a large sample $(1-\delta)100\%$ prediction region, but regions with smaller volume may exist.

The nonparametric prediction region for multivariate linear regression of Theorem 3 uses $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ in (5), and has simple geometry. Let R_r be the nonparametric prediction region applied to the residuals $\hat{\boldsymbol{\epsilon}}_i$. Then R_r is a hyperellipsoid with center $\mathbf{0}$, and the nonparametric prediction region is the hyperellipsoid R_r translated to have center $\hat{\boldsymbol{y}}_f$. See examples in section 5.

4 Testing Hypotheses

Consider testing a linear hypothesis $H_0 : \boldsymbol{LB} = \mathbf{0}$ versus $H_1 : \boldsymbol{LB} \neq \mathbf{0}$ where $\boldsymbol{L}$ is a full rank $r \times p$ matrix. Let $\boldsymbol{H} = \hat{\boldsymbol{B}}^T \boldsymbol{L}^T [\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}$. Let the error or residual sum of squares and cross products matrix be

$$\boldsymbol{W}_e = \hat{\boldsymbol{E}}^T\hat{\boldsymbol{E}} = (\boldsymbol{Z} - \hat{\boldsymbol{Z}})^T(\boldsymbol{Z} - \hat{\boldsymbol{Z}}) = \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{X}\hat{\boldsymbol{B}} = \boldsymbol{Z}^T[\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Z}.$$

Then $\boldsymbol{W}_e/(n-p) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the ordered eigenvalues of $\boldsymbol{W}_e^{-1}\boldsymbol{H}$. Then there are four commonly used test statistics.

The Roy's maximum root statistic is $\lambda_{max}(\boldsymbol{L}) = \lambda_1$.

The Wilks' Λ statistic is $\Lambda(\boldsymbol{L}) = |(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{W}_e| = |\boldsymbol{W}_e^{-1}\boldsymbol{H} + \boldsymbol{I}|^{-1} = \prod_{i=1}^{m}(1+\lambda_i)^{-1}$.

The Pillai's trace statistic is $V(\boldsymbol{L}) = tr[(\boldsymbol{H} + \boldsymbol{W}_e)^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \frac{\lambda_i}{1+\lambda_i}$.

The Hotelling-Lawley trace statistic is $U(\boldsymbol{L}) = tr[\boldsymbol{W}_e^{-1}\boldsymbol{H}] = \sum_{i=1}^{m} \lambda_i$.

Theorem 4. The Hotelling-Lawley trace statistic

$$U(\boldsymbol{L}) = \frac{1}{n-p}[vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})]. \quad (7)$$

The proof of Theorem 4 is in the appendix. Some notation is useful to show (7) and to show that $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$ under mild conditions if H_0 is true. Following Henderson and Searle (1979), let matrix $\boldsymbol{A} = [\boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \dots \quad \boldsymbol{a}_p]$. Then the vec operator stacks the columns of $\boldsymbol{A}$ on top of one another, and $\boldsymbol{A} \otimes \boldsymbol{B}$ is the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$. An important fact is that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are nonsingular square matrices, then $[\boldsymbol{A} \otimes \boldsymbol{B}]^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$. The following assumption is important.

Assumption D1: Let h_i be the ith diagonal element of $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Assume $\max_{1\le i\le n} h_i \to 0$ as $n \to \infty$, assume that the zero mean iid errors have finite fourth moments, and assume that $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \xrightarrow{P} \boldsymbol{W}^{-1}$.

Then for the least squares estimator, Su and Cook (2012) show that if assumption D1 holds, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$ is $\sqrt{n}$ consistent and $\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$.

Theorem 5. If assumption D1 holds and if H_0 is true, then $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$.

The proof of Theorem 5 is in the appendix. Kakizawa (2009) shows, under stronger assumptions than Theorem 5, that for a large class of iid error distributions, the following test statistics have the same χ^2_{rm} limiting distribution when H_0 is true, and the same noncentral $\chi^2_{rm}(\omega^2)$ limiting distribution with noncentrality parameter ω^2 when H_0 is false under a local alternative. Hence the three tests are robust to the assumption of normality. The limiting null distribution is well known when the zero mean errors

are iid from a multivariate normal distribution. See Khattree and Naik (1999, p. 68): $(n-p)U(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, $(n-p)V(\boldsymbol{L}) \xrightarrow{D} \chi^2_{rm}$, and $-[n-p-0.5(m-r+3)]\log(\Lambda(\boldsymbol{L})) \xrightarrow{D} \chi^2_{rm}$. Results from Kshirsagar (1972, p. 301) suggest that the chi-square approximation is very good if $n \geq 3(m^2+p^2)$ for multivariate normal errors.

Theorems 4 and 5 are useful for relating multivariate tests with the partial F test for multiple linear regression that tests whether a reduced model that omits some of the predictors can be used instead of the full model that uses all p predictors. The partial F test statistic is

$$F_R = \left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F)$$

where the residual sums of squares $SSE(F)$ and $SSE(R)$ and degrees of freedom df_F and df_r are for the full and reduced model while the mean square error $MSE(F)$ is for the full model. Let the null hypothesis for the partial F test be $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{0}$ where $\boldsymbol{L}$ sets the coefficients of the predictors in the full model but not in the reduced model to 0. Seber and Lee (2003, p. 100) shows that

$$F_R = \frac{[\boldsymbol{L}\hat{\boldsymbol{\beta}}]^T(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}[\boldsymbol{L}\hat{\boldsymbol{\beta}}]}{r\hat{\sigma}^2}$$

is distributed as $F_{r,n-p}$ if H_0 is true and the errors are iid $N(0,\sigma^2)$. Note that for multiple linear regression with $m = 1$, $F_R = (n-p)U(\boldsymbol{L})/r$ since $\hat{\boldsymbol{\Sigma}}_\epsilon^{-1} = 1/\hat{\sigma}^2$. Hence the scaled Hotelling Lawley test statistic is the partial F test statistic extended to $m > 1$ predictor variables by Theorem 4.

By Theorem 5, for example, $rF_R \xrightarrow{D} \chi^2_r$ for a large class of nonnormal error distribution. If $Z_n \sim F_{k,d_n}$, then $Z_n \xrightarrow{D} \chi^2_k/k$ as $d_n \to \infty$. Hence using the $F_{r,n-p}$ approximation gives a large sample test with correct asymptotic level, and the partial F test is robust

to nonnormality.

Similarly, using an $F_{rm,n-pm}$ approximation for the following test statistics gives large sample tests with correct asymptotic level by Kakizawa (2009) and similar power for large n. The large sample test will have correct asymptotic level as long as the denominator degrees of freedom $d_n \to \infty$ as $n \to \infty$, and $d_n = n - pm$ reduces to the partial F test if $m = 1$ and $U(\boldsymbol{L})$ is used. Then the three test statistics are

$$\frac{-[n-p-0.5(m-r+3)]}{rm} \ \log(\Lambda(\boldsymbol{L}), \ \ \frac{n-p}{rm} \ V(\boldsymbol{L}), \text{ and } \frac{\mathrm{n-p}}{\mathrm{rm}} \ \mathrm{U}(\boldsymbol{L}).$$

Following Khattree and Naik (1999, p. 67) for the Roy's largest root test, if $h = \max(r, m)$, use

$$\frac{n-p-h+r}{h}\lambda_{max}(\boldsymbol{L}) \approx F(h, n-p-h+r).$$

The simulations in section 5 suggest that this approximation is good for $r = 1$ but poor for $r > 1$. Anderson (1984, p. 333) states that Roy's largest root test has the greatest power if $r = 1$ but is an inferior test for $r > 1$.

Multivariate analogs of tests for multiple linear regression can be derived with appropriate choice of $\boldsymbol{L}$. Assume a constant $x_1 = 1$ is in the model. The analog of the ANOVA F test for multiple linear regression is the MANOVA F test that uses $\boldsymbol{L} = [\boldsymbol{0} \ \ \boldsymbol{I}_{p-1}]$ to test whether the nontrivial predictors are needed in the model. This test should reject H_0 if the response and residual plots look good, n is large enough and at least one response plot does not look like the corresponding residual plot. A response plot for Y_j will look like a residual plot if the identity line appears almost horizontal, hence the range of $\hat{Y}_j$ is small.

The F_j test of hypotheses uses $\boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$, where the 1 is in the jth

position, to test whether the jth predictor is needed in the model given that the other $p-1$ predictors are in the model. This test is an analog of the t tests for multiple linear regression.

The MANOVA partial F test is used to test whether a reduced model is good where the reduced model deletes r of the variables from the full model. For this test, the ith row of $\boldsymbol{L}$ has a 1 in the position corresponding to the ith variable to be deleted. Omitting the jth variable corresponds to the F_j test while omitting variables $x_2, ..., x_p$ corresponds to the MANOVA F test. Using $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_k]$ tests whether the last k predictors are needed in the multivariate linear regression model given that the remaining predictors are in the model.

5 Examples and Simulations

The semiparametric and parametric MVN prediction regions applied to the $\hat{\boldsymbol{z}}_i$ are only conjectured to be large sample prediction regions, but are added to the DD plot as visual aids. Cases below the horizontal line that crosses the identity line correspond to the semiparametric region while cases below the horizontal line that ends at the identity line correspond to the parametric MVN region. A vertical line dropped down from this point of intersection does correspond to a large sample prediction region for multivariate normal data. Note that $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$, and adding a constant $\hat{\boldsymbol{y}}_f$ to all of the residual vectors does not change the Mahalanobis distances, so the DD plot of the residuals can be used to display the prediction regions.

Example 1. Cook and Weisberg (1999a, p. 351, 433, 447) gives a data set on 82 mussels sampled off the coast of New Zealand. Let $Y_1 = \log(S)$ and $Y_2 = \log(M)$ where S is the shell mass and M is the muscle mass. The predictors are $X_2 = L$, $X_3 = \log(W)$ and $X_4 = H$: the shell length, log(width) and height. Figures 1 and 2 give the response and residual plots for Y_1 and Y_2. The response plots show strong linear relationships. For Y_1, case 79 sticks out while for Y_2, cases 8, 25 and 48 are not fit well. Highlighted cases had Cook's distance $> \min(0.5, 2p/n)$. See Cook (1977). Figure 3 shows the DD plot of the residual vectors. The plotted points are highly correlated but do not cover the identity line, suggesting an elliptically contoured error distribution that is not multivariate normal. The nonparametric 90% prediction region for the residuals consists of the points to the left of the vertical line $MD = 2.60$. Cases 8, 48 and 79 have especially large distances. The four Hotelling Lawley F_j statistics were greater than 5.77 with pvalues less than 0.005, and the MANOVA F statistic was 337.8 with pvalue ≈ 0.

The response, residual and DD plots are effective for finding influential cases, for checking linearity and whether the error distribution is multivariate normal or some other elliptically contoured distribution, and for displaying the nonparametric prediction region. Note that cases to the right of the vertical line correspond to cases that are not in their prediction region. These are the cases corresponding to residual vectors with large Mahalanobis distances. Adding a constant does not change the distance, so the DD plot for the residuals is the same as the DD plot for the $\hat{\boldsymbol{z}}_i$.

Suppose the same model is used except $Y_2 = M$. Then the response and residual plots for Y_1 remain the same, but the plots shown in Figure 4 show curvature about the identity and $r = 0$ lines. Hence the linearity condition is violated. Figure 5 shows that the

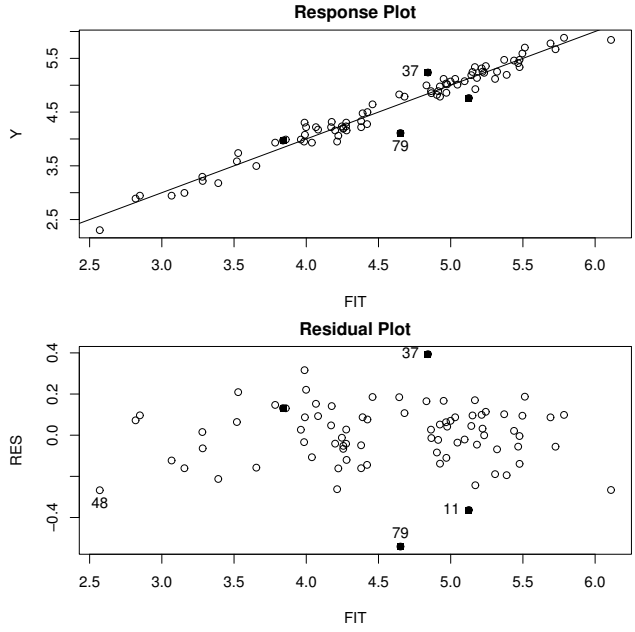

Figure 1: Plots for $Y_1 = \log(S)$.

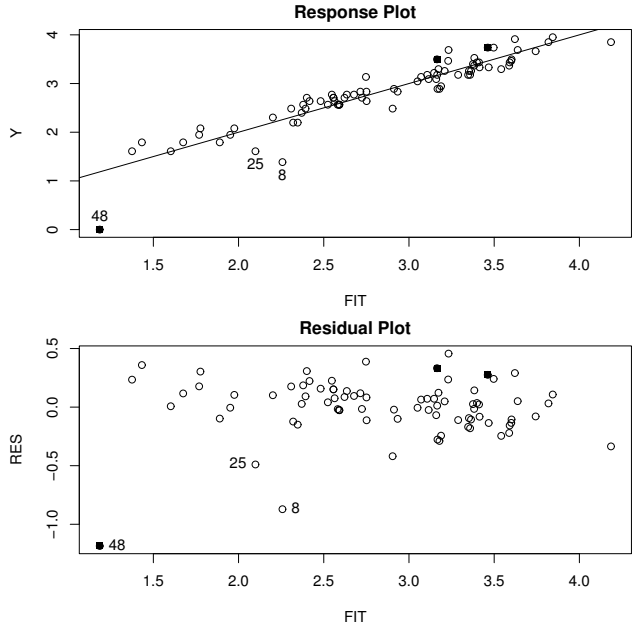

Figure 2: Plots for $Y_2 = \log(M)$.

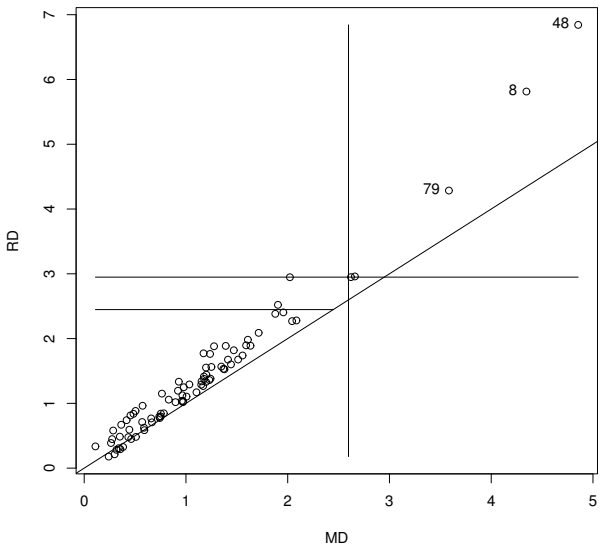

Figure 3: DD Plot of the Residual Vectors for the Mussel Data.

plotted points in the DD plot have correlation well less than one, suggesting that the error distribution is no longer elliptically contoured. The nonparametric 90% prediction region for the residual vectors consists of the points to the left of the vertical line $MD = 2.52$, and contains 95% of the data. Note that the plots can be used to quickly assess whether power transformations have resulted in a linear model, and whether influential cases are present.

Example 2. Consider the one way MANOVA model on the famous iris data set with $n = 150$ and $p = 3$ species of iris: setosa, versicolor, and virginica. The $m = 4$ variables are $Y_1 =$ *sepal length*, $Y_2 =$ *sepal width*, $Y_3 =$ *petal length* and $Y_4 =$ *petal width*. See Becker, Chambers and Wilks (1988). Figure 6 shows the response and residual plots for Y_4. Note that the spread of the three dot plots is similar. The dot plot intersects the identity line at the sample mean of the cases in the dot plot. The setosa cases in

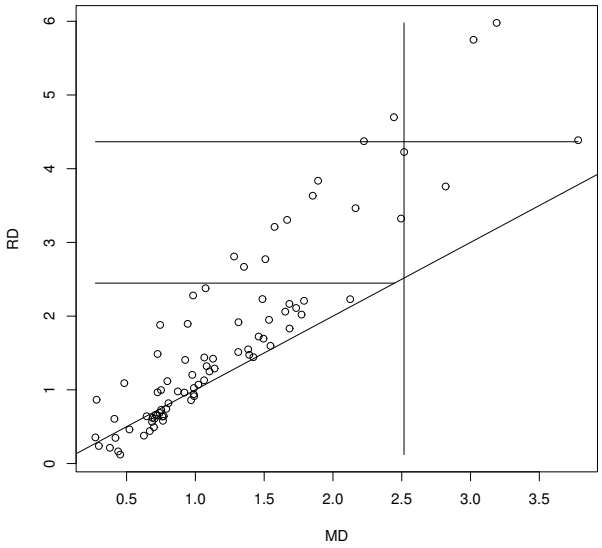

Figure 4: Plot for $Y_2 = M$.

lowest dot plot have a sample mean of 0.246 and the horizontal line $Y_4 = 0.246$ is below the dot plots for versicolor and virginica which have means of 1.326 and 2.026. Hence the mean petal widths differ for the three species, and it is easier to see this difference in the response plot than the residual plot. The plots for the other three variables are similar. Figure 7 shows that the DD plot suggests that the error distribution is elliptically contoured but not multivariate normal. For the one way MANOVA model, a prediction region for $\boldsymbol{y}_f$ would only be valid for an $\boldsymbol{x}_f$ which was observed, i.e., for $\boldsymbol{x}_f = \boldsymbol{x}_j$, since only observed values of the categorical predictor variables make sense. The 90% nonparametric prediction region corresponds to $\boldsymbol{y}$ with distances to the left of the vertical line $MD = 3.2$.

A small simulation was used to study the prediction region and the Wilks' Lambda test, the Pillai's trace test, the Hotelling Lawley trace test, and the Roy's largest root

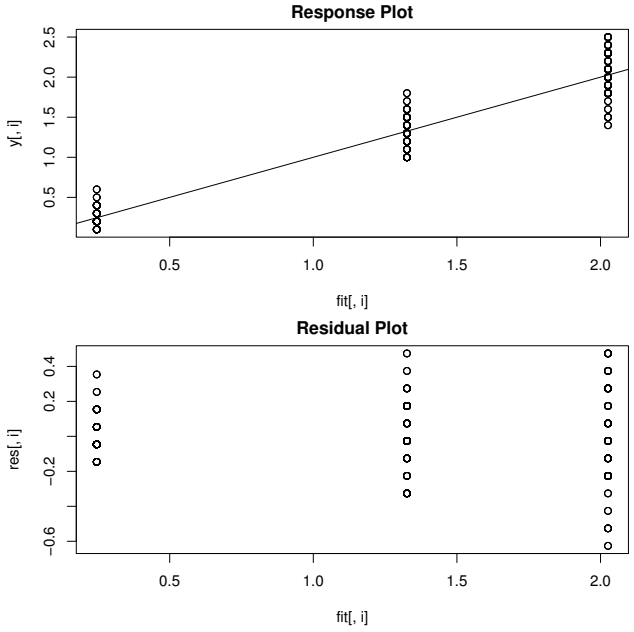

Figure 5: Plots for Y_4 = Petal Width.

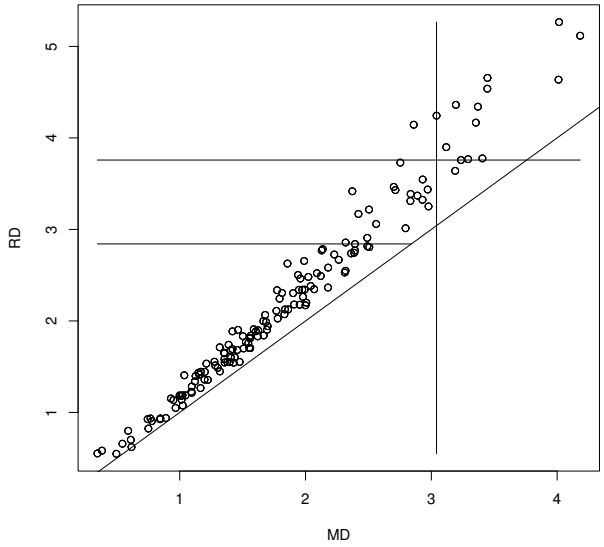

Figure 6: DD Plot of the Residual Vectors for Iris Data.

test for the F_j tests and the MANOVA F test for multivariate linear regression. The first row of $\boldsymbol{B}$ was always $\mathbf{1}^T$ and the last row of $\boldsymbol{B}$ was always $\mathbf{0}^T$. When the null hypothesis for the MANOVA F test is true, all but the first row corresponding to the constant are equal to $\mathbf{0}^T$. When $p \geq 3$ and the null hypothesis for the MANOVA F test is false, then the second to last row of $\boldsymbol{B}$ is (1, 0, ..., 0), the third to last row is (1, 1, 0, ..., 0) et cetera as long as the first row is not changed from $\mathbf{1}^T$. First $m \times 1$ error vectors $\boldsymbol{w}_i$ were generated such that the m errors are iid with variance σ^2. Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then $\boldsymbol{\epsilon}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}} = \sigma^2 \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = \sigma^2[1 + (m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = \sigma^2[2\psi + (m-2)\psi^2]$ where $\psi = 0.10$. Hence the correlations are $(2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$. As ψ gets close to 1, the data clusters about the line in the direction of $(1, ..., 1)^T$. Used $\boldsymbol{w}_i \sim N_m(\mathbf{0}, \boldsymbol{I}), \boldsymbol{w}_i \sim (1-\tau)N_m(\mathbf{0}, \boldsymbol{I}) + \tau N_m(\mathbf{0}, 25\boldsymbol{I})$ with $0 < \tau < 1$ and $\tau = 0.25$ in the simulation, $\boldsymbol{w}_i \sim$ multivariate t_d with $d = 7$ degrees of freedom, or $\boldsymbol{w}_i \sim$ lognormal - E(lognormal): where the m components of $\boldsymbol{w}_i$ were iid with distribution $e^z - E(e^z)$ where $z \sim N(0, 1)$. Only the lognormal distribution is not elliptically contoured.

The simulation used 5000 runs, and H_0 was rejected if the F statistic was greater than $F_{d_1,d_2}(0.95)$ where $P(F_{d_1,d_2} < F_{d_1,d_2}(0.95)) = 0.95$ with $d_1 = rm$ and $d_2 = n - mp$ for the test statistics

$$\frac{-[n-p-0.5(m-r+3)]}{rm} \log(\Lambda(\boldsymbol{L}), \quad \frac{n-p}{rm} V(\boldsymbol{L}), \text{ and } \frac{\mathrm{n-p}}{\mathrm{rm}} \mathrm{U}(\boldsymbol{L})$$

while $d_1 = h = \max(r, m)$ and $d_2 = n - p - h + r$ for the test statistic

$$\frac{n-p-h+r}{h} \lambda_{max}(\boldsymbol{L}).$$

Denote these statistics by W, P, HL and R. Let the coverage be the proportion of times that H_0 is rejected. Want coverage near 0.05 when H_0 is true and coverage close to 1 for good power when H_0 is false. With 5000 runs, coverage outside of (0.04,0.06) suggests that the true coverage is not 0.05. Coverages are tabled for the F_1, F_2, F_{p-1}, and F_p test and for the MANOVA F test denoted by F_M. The null hypothesis H_0 was always true for the F_p test and always false for the F_1 test. When the MANOVA F test was true, H_0 was true for the F_j tests with $j \neq 1$. When the MANOVA F test was false, H_0 was false for the F_j tests with $j \neq p$, but the F_{p-1} test should be hardest to reject for $j \neq p$ by construction of $\boldsymbol{B}$ and the error vectors.

When the null hypothesis H_0 was true, simulated values started to get close to nominal levels for $n \geq 0.75(m+p)^2$, and were fairly good for $n \geq 1.5(m+p)^2$. The exception was Roy's test which rejects H_0 far too often if $r > 1$. See Table 1 where want values for the F_1 test to be close to 1 since H_0 is false for the F_1 test and want values close to 0.05, otherwise. Roy's test was very good for the F_j tests but very poor for the MANOVA F test. Results are shown for $m = p = 10$. Results from Berndt and Savin (1977) suggest that Pillai's test will reject H_0 less often than Wilks' test which will reject less often than the Hotelling Lawley test.

In Table 2, H_0 is only true for the F_p test where $p = m$, and want values in the F_p column near 0.05. Want values near 1 for high power otherwise. If H_0 is false, often H_0 will be rejected for small n. For example, if $n \geq 10p$, then the response and residual plots should start to look good, and the MANOVA F test should be rejected if there is at least one response variable where the identity line fits the plotted points better than any horizontal line. For the simulated data, had fair power for n not much larger than

Table 1: Test Coverages: MANOVA F H_0 is True.

$\boldsymbol{w}$ dist	n	test	F_1	F_2	F_{p-1}	F_p	F_M
MVN	300	W	1	0.043	0.042	0.041	0.018
MVN	300	P	1	0.040	0.038	0.038	0.007
MVN	300	HL	1	0.059	0.058	0.057	0.045
MVN	300	R	1	0.051	0.049	0.048	0.993
MVN	600	W	1	0.048	0.043	0.043	0.034
MVN	600	P	1	0.046	0.042	0.041	0.026
MVN	600	HL	1	0.055	0.052	0.050	0.052
MVN	600	R	1	0.052	0.048	0.047	0.994
MIX	300	W	1	0.042	0.043	0.044	0.017
MIX	300	P	1	0.039	0.040	0.042	0.008
MIX	300	HL	1	0.057	0.059	0.058	0.039
MIX	300	R	1	0.050	0.050	0.051	0.993
MVT(7)	300	W	1	0.048	0.036	0.045	0.020
MVT(7)	300	P	1	0.046	0.032	0.042	0.011
MVT(7)	300	HL	1	0.064	0.049	0.058	0.045
MVT(7)	300	R	1	0.055	0.043	0.051	0.993
LN	300	W	1	0.043	0.047	0.040	0.020
LN	300	P	1	0.039	0.045	0.037	0.009
LN	300	HL	1	0.057	0.061	0.058	0.041
LN	300	R	1	0.049	0.055	0.050	0.994

mp. Results are shown for the lognormal distribution.

The same type of data and 5000 runs were used to simulate the prediction regions for $\boldsymbol{y}_f$ given $\boldsymbol{x}_f$ for multivariate regression. With n=100, m=2, and p=4, the nominal coverage of the prediction region is 90%, and 92% of the training data is covered. Following Olive (2013), consider the prediction region $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. Then the ratio of the prediction region volumes

$$\frac{h_i^m \sqrt{det(\boldsymbol{C}_i)}}{h_2^m \sqrt{det(\boldsymbol{C}_2)}}$$

was recorded where $i = 1$ was the nonparametric region, $i = 2$ was the semiparametric region, and $i = 3$ was the parametric MVN region. Here h_1 and h_2 were the cutoff $D_{(U_n)}(T_i, \boldsymbol{C}_i)$ for $i = 1, 2$, and $h_3 = \sqrt{\chi^2_{m,q_n}}$.

If, as conjectured, the RMVN estimator is a consistent estimator when applied to the residual vectors instead of iid data, then the volume ratios converge in probability to 1 if the iid zero mean errors $\sim N_m(\boldsymbol{0}, \boldsymbol{\Sigma_\epsilon})$, and the volume ratio converges to 1 for $i = 1$ for a large class of elliptically contoured distributions. These volume ratios were denoted by voln and volm for the nonparametric and parametric MVN regions. The coverage was the proportion of times the prediction region contained $\boldsymbol{y}_f$ where ncov, scov and mcov are for the nonparametric, semiparametric and parametric MVN regions.

In the simulations, took $n = 3(m + p)^2$ and $m = p$. Table 3 shows that the coverage of the nonparametric region was close to 0.9 in all cases. The volume ratio voln was fairly close to 1 for the three elliptically contoured distributions. Since the volume of the prediction region is proportional to h^m, the volume can be very small if h is too small and m is large. Parametric prediction regions usually give poor estimates of h

Table 2: Test Coverages: MANOVA F H_0 is False.

n	$m = p$	test	F_1	F_2	F_{p-1}	F_p	F_M
30	5	W	0.012	0.222	0.058	0.000	0.006
30	5	P	0.000	0.000	0.000	0.000	0.000
30	5	HL	0.382	0.694	0.322	0.007	0.579
30	5	R	0.799	0.871	0.549	0.047	0.997
50	5	W	0.984	0.955	0.644	0.017	0.963
50	5	P	0.971	0.940	0.598	0.012	0.871
50	5	HL	0.997	0.979	0.756	0.053	0.991
50	5	R	0.996	0.978	0.744	0.049	1
105	10	W	0.650	0.970	0.191	0.000	0.633
105	10	P	0.109	0.812	0.050	0.000	0.000
105	10	HL	0.964	0.997	0.428	0.000	1
105	10	R	1	1	0.892	0.052	1
150	10	W	1	1	0.948	0.032	1
150	10	P	1	1	0.941	0.025	1
150	10	HL	1	1	0.966	0.060	1
150	10	R	1	1	0.965	0.057	1
450	20	W	1	1	0.999	0.020	1
450	20	P	1	1	0.999	0.016	1
450	20	HL	1	1	0.999	0.035	1
450	20	R	1	1	0.999	0.056	1

Table 3: Coverages for 90% Prediction Regions.

$\boldsymbol{w}$ dist	n	$m = p$	ncov	scov	mcov	nvol	mvol
MVN	48	2	0.901	0.905	0.888	0.941	0.964
MVN	300	5	0.889	0.887	0.890	1.006	1.015
MVN	1200	10	0.899	0.896	0.896	1.004	1.001
MIX	48	2	0.912	0.927	0.710	0.872	0.097
MIX	300	5	0.906	0.911	0.680	0.882	0.001
MIX	1200	10	0.904	0.911	0.673	0.889	0+
MVT(7)	48	2	0.903	0.910	0.825	0.914	0.646
MVT(7)	300	5	0.899	0.909	0.778	0.916	0.295
MVT(7)	1200	10	0.906	0.911	0.726	0.919	0.061
LN	48	2	0.912	0.926	0.651	0.729	0.090
LN	300	5	0.915	0.917	0.593	0.696	0.009
LN	1200	10	0.912	0.916	0.593	0.679	0+

when the parametric distribution is misspecified. Hence the parametric MVN region only performed well for multivariate normal data.

6 Conclusions

Multivariate linear regression is a semiparametric method that is nearly as easy to use as multiple linear regression if m is small. The m response and residual plots should be made as well as the DD plot, and the response and residual plots are very useful for the $m = 1$ case of multiple linear regression and experimental design. These plots speed up the model building process for multivariate linear models since the success of power transformations achieving linearity can be quickly assessed and influential cases can be quickly detected. Work is needed on variable selection and on determining the sample sizes for when the tests and prediction regions start to work well. Response and residual plots can look good for $n \geq 10p$, but for testing and prediction regions, may need $n \geq k(m+p)^2$ where $0.5 \leq k \leq 3$ even for well behaved elliptically contoured error distributions.

The *R* software was used to make plots and software. See R Development Core Team (2011). The function `mpredsim` was used to simulate the prediction regions, `mregsim` was used to simulate the tests of hypotheses, and `mregddsim` simulated the DD plots for various distributions. The function `mltreg` makes the response and residual plots and computes the F_j, MANOVA F and MANOVA partial F test pvalues while the function `ddplot4` makes the DD plots.

Appendix

Proof of Theorem 1. Let B_n denote the subset of the sample space on which both $\hat{\boldsymbol{\Sigma}}_{1,n}$ and $\hat{\boldsymbol{\Sigma}}_{2,n}$ have inverses. Then $P(B_n) \to 1$ as $n \to \infty$. Now

$$D^2_{\boldsymbol{x}}(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) =$$

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \; - \; \frac{\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{-\boldsymbol{\Sigma}^{-1}}{a_j} + \hat{\boldsymbol{\Sigma}}_j^{-1} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) +$$

$$(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) = \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T (-\boldsymbol{\Sigma}^{-1} + a_j \, \hat{\boldsymbol{\Sigma}}_j^{-1}) (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) \; +$$

$$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \left(\frac{\boldsymbol{\Sigma}^{-1}}{a_j} \right) (\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)$$

$$= \frac{1}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

$$+ \frac{2}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j) + \frac{1}{a_j} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j) + \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T [a_j \hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)$$

on B_n, and the last three terms are $o_P(1)$ under a) and $O_P(n^{-\delta})$ under b).

Proof of Theorem 2. a) Suppose $(\boldsymbol{x}_f, \boldsymbol{y}_f) = (\boldsymbol{x}_i, \boldsymbol{y}_i)$. Then

$$D^2_{\boldsymbol{y}_i}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) = (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) = \hat{\boldsymbol{\epsilon}}_i^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \hat{\boldsymbol{\epsilon}}_i = D^2_{\hat{\boldsymbol{\epsilon}}_i}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}).$$

Hence $\boldsymbol{y}_i$ is in the ith prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}(\hat{\boldsymbol{y}}_i, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ iff $\hat{\boldsymbol{\epsilon}}_i$ is in prediction region $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{(U_n)}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$, but exactly U_n of the $\hat{\boldsymbol{\epsilon}}_i$ are in the latter region by construction, if $D_{(U_n)}$ is unique. Since $D_{(U_n)}$ is the $100(1-\delta)$th percentile of the D_i asymptotically, $U_n/n \to 1 - \delta$.

b) Let $P[D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})] = 1 - \delta$. Since $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, Theorem 1 shows that if $(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} (E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ then $D(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Hence the percentiles of the distances also converge in probability, and the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\}$ converges to $1 - \delta =$ the probability that $\boldsymbol{y}_f$ is in $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$.

c) The asymptotically optimal prediction region is the region with the smallest volume (hence highest density) such that the coverage is $1-\delta$, as $n \to \infty$. This region is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ if the asymptotically optimal region for the $\boldsymbol{\epsilon}_i$ is $\{\boldsymbol{z} : D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \leq D_{1-\delta}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$. Hence the result follows by b).

Proof of Theorem 3. Multivariate linear regression with least squares satisfies Theorem 2 by Su and Cook (2012). Let $(T, \boldsymbol{C})$ be the sample mean and sample covariance matrix (1) applied to the $\hat{\boldsymbol{z}}_i$. The sample mean and sample covariance matrix of the residual vectors is $(\boldsymbol{0}, \boldsymbol{S}_r)$ since least squares was used. Hence the $\hat{\boldsymbol{z}}_i = \hat{\boldsymbol{y}}_f + \hat{\boldsymbol{\epsilon}}_i$ have sample covariance matrix $\boldsymbol{S}_r$, and sample mean $\hat{\boldsymbol{y}}_f$. Hence $(T, \boldsymbol{C}) = (\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$, and the $D_i(\hat{\boldsymbol{y}}_f, \boldsymbol{S}_r)$ are used to compute $D_{(U_n)}$.

Proof of Theorem 4. Using the Searle (1982, p. 333) identity

$tr(\boldsymbol{A}\boldsymbol{G}^T\boldsymbol{D}\boldsymbol{G}\boldsymbol{C}) = [vec(\boldsymbol{G})]^T[\boldsymbol{C}\boldsymbol{A} \otimes \boldsymbol{D}^T][vec(\boldsymbol{G})]$, it follows that

$(n-p)U(\boldsymbol{L}) = tr[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}\hat{\boldsymbol{B}}^T\boldsymbol{L}^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{B}}]$

$= [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] = T$ where $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1}, \boldsymbol{G} = \boldsymbol{L}\hat{\boldsymbol{B}}, \boldsymbol{D} = [\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}$, and $\boldsymbol{C} = \boldsymbol{I}$. Hence (7) holds.

Proof of Theorem 5. By Su and Cook (2012), $\sqrt{n}\ \ vec(\hat{\boldsymbol{B}} - \boldsymbol{B}) \xrightarrow{D} N_{pm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{W})$. Then under H_0, $\sqrt{n}\ \ vec(\boldsymbol{L}\hat{\boldsymbol{B}}) \xrightarrow{D} N_{rm}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \otimes \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)$, and

$n\ \ [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}$. This result also holds if $\boldsymbol{W}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ are replaced by $\hat{\boldsymbol{W}} = n(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}$. Hence under H_0 and using the proof of Theorem 4, $T = (n-p)U(\boldsymbol{L}) = [vec(\boldsymbol{L}\hat{\boldsymbol{B}})]^T[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}^{-1} \otimes (\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}][vec(\boldsymbol{L}\hat{\boldsymbol{B}})] \xrightarrow{D} \chi^2_{rm}$.

References

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis,* 2nd ed.,

Wiley, New York, NY.

Becker, R.A., Chambers, J.M., and Wilks, A.R. (1988), *The New S Language a Programming Environment for Data Analysis and Graphics,* Wads-worth and Brooks/Cole, Pacific Grove, CA.

Berndt, E.R. and Savin, N.E. (1977), "Conflict Among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model," *Econometrika*, 45, 1263-1277.

Cook, R.D. (1977), "Deletion of Influential Observations in Linear Regression," *Technometrics,* 19, 15-18.

Cook, R.D., and Weisberg, S. (1999a), *Applied Regression Including Computing and Graphics,* Wiley, New York, NY.

Cook, R.D., and Weisberg, S. (1999b), "Graphs in Statistical Analysis: is the Medium the Message?" *The American Statistician,* 53, 29-37.

Cook, R.D., and Olive, D.J. (2001), "A Note on Visualizing Response Transformations in Regression," *Technometrics,* 43, 443-449.

Henderson, H.V., and Searle, S.R. (1977), "Vec and Vech Operators for Matrices, with Some Uses in Jacobians and Multivariate Statistics," *The Canadian Journal of Statistics,* 7, 65-81.

Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

Kakizawa, Y. (2009), "Third-Order Power Comparisons for a Class of Tests for Multivariate Linear Hypothesis Under General Distributions," *Journal of Multivariate Analysis*, 100, 473-496.

Khattree, R., and Naik, D.N. (1999), *Applied Multivariate Statistics with SAS Software,*

2nd ed., SAS Institute, Cary, NC.

Kshirsagar, A.M. (1972), *Multivariate Analysis*, Marcel Dekker, New York, NY.

Olive, D.J. (2002), "Applications of Robust Distances for Regression," *Technometrics,* 44, 64-71.

Olive, D.J. (2007), "Prediction Intervals for Regression Models," *Computational Statistics and Data Analysis,* 51, 3115-3122.

Olive, D.J. (2013), "Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data," *International Journal of Statistics and Probability*, 2, 90-100.

Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," Preprint, see (www.math.siu.edu/olive/preprints.htm).

R Development Core Team (2011), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

Searle, S.R. (1982), *Matrix Algebra Useful for Statistics*, Wiley, New York, NY.

Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., Wiley, New York, NY.

Su, Z., and Cook, R.D. (2012), "Inner Envelopes: Efficient Estimation in Multivariate Linear Regression," *Biometrika*, 99, 687-702.

Zhang, J., Olive, D.J., and Ye, P. (2012), "Robust Covariance Matrix Estimation With Canonical Correlation Analysis," *International Journal of Statistics and Probability*, 1, 119-136.